PROJETO APLICADO III

SISTEMA DE RECOMENDAÇÃO DE LIVROS PARA APOIAR O ODS 9: INDÚSTRIA, INOVAÇÃO E INFRAESTRUTURA

Gustavo José Fermiano – RA: 10440929

Kelly Haro Vasconcellos – RA: 10441014

Wesley Rodrigo dos Santos – RA: 10433408

Coleta de dados

- Para este estudo, foi utilizado o conjunto de dados público "Book-Crossing Dataset" (https://www.kaggle.com/datasets/ruchi798/bookcrossing-dataset/data), disponível na plataforma Kaggle.
- BX-Book-Ratings.csv: Contém o registro das avaliações dos livros, sendo as colunas principais: User-ID (identificador do usuário), ISBN (identificador do livro) e Book-Rating (nota atribuída).
- BX-Books.csv: Funciona como um catálogo de metadados, contendo informações dos livros, como ISBN, Book-Title (título) e Book-Author (autor).

Pré-processamento e Limpeza dos Dados

- Remoção de avaliações com nota "0".
- Filtragem de usuários (Número mínimo de avaliações)
- Filtragem de livros (Popularidade mínima)
- Redução de aproximadamente 75% das avaliações do dataset.

Seleção do Algoritmo

- Modelo de filtragem colaborativa baseada em usuários.
- Algoritmo selecionado: k-Nearest Neighbors (KNN)
- Métrica de similaridade: Similaridade do Cosseno
- Parâmetro k: ajustado empiricamente, k=30.
- Ambiente utilizado: Google Colab (Jupyter Notebook)
- Linguagem utilizada: Python

Funções

- gerar_recomendacoes_por_id: retorna um número pré-definido (padrão: 10) de recomendações de livros de acordo com o User-ID (identificador do usuário) de um usuário específico
- novo_usuario: gera um novo usuário criado com base em avaliações de 3 livros dentre os 500 mais populares da base, chamando a função gerar_recomendacoes_por_id para retornar um número pré-definido de recomendações para o usuário recém-criado
- numero_recomendacoes_por_id: retorna o número de recomendações gerados por User-ID (identificador do usuário). Usado para ajustar os parâmetros do algoritmo.

Métricas de avaliação

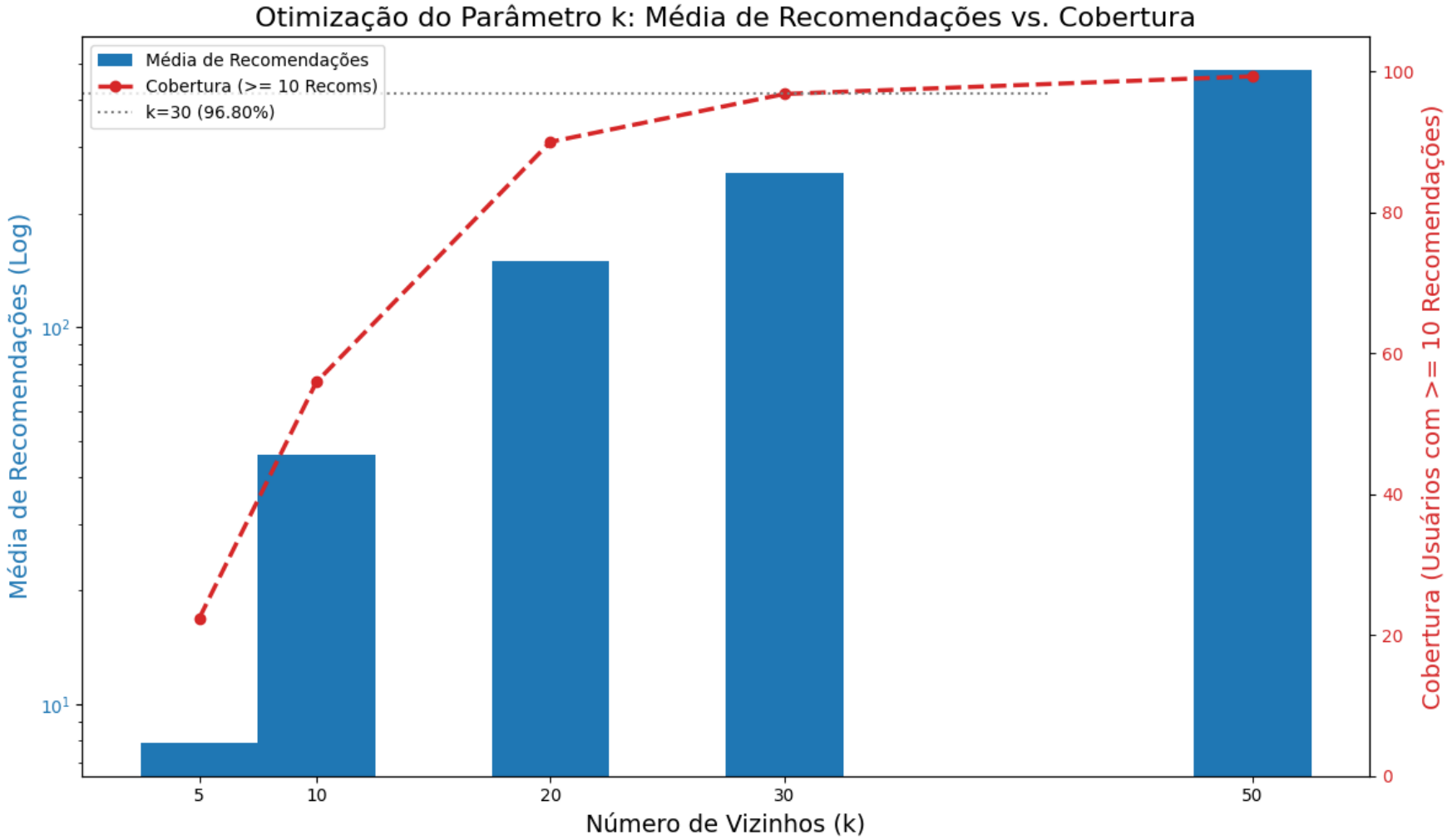
- Métricas de avaliação não adequadas a um sistema de recomendação, como RMSE e MAE
- Sistema não é do tipo regressão, ou seja, não é utilizado para prever notas.
- Dados não rotulados, e dificuldade de rotulação dos dados para um sistema de recomendação, dificultaram a aplicação de métricas de avaliação convencionais.

Otimização e ajustes

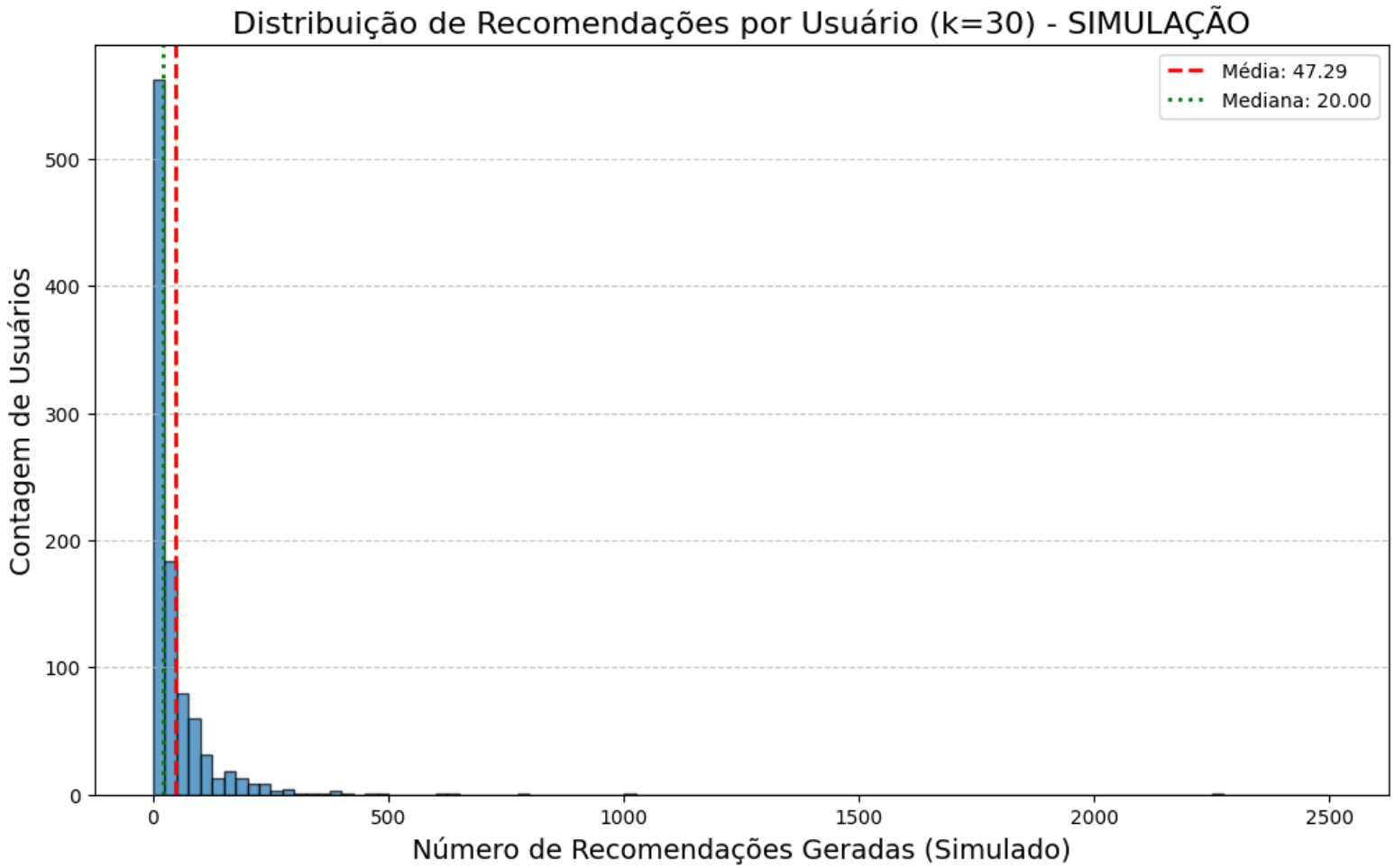
- Realizados testes com diferentes parâmetros de k (número de vizinhos) para ajustar sistema de recomendação.
- Especificidade vs. generalização

Número de vizinhos (k)	Média de recomendações	Mínimo de recomendaçõe s	Máximo de recomendaçõe s	Porcentagem de recomendações maior ou igual a 10
5	7,91	0	218	22,40%
10	45,71	0	1884	55,90%
20	149,24	0	2082	90,00%
30	256,48	0	2258	96,80%
50	481,67	0	2390	99,30%

Ajuste do parâmetro K



Ajuste do parâmetro K



Limitações

- Usuários com zero recomendações
- Usuários com número excessivo de recomendações
- Alta variância no número de recomendações
- Problema de cold start (e solução parcial)
- Dificuldade na avaliação do modelo
- Escalabilidade e custo computacional

Trabalhos futuros: Sugestões de melhorias

- Implementação de sistema híbrido, combinado com filtragem baseada em conteúdo
- Algoritmos alternativos, como uso de redes neurais
- Métricas alternativas de similaridade
- Implantação de testes e validações adicionais, como métricas de ranking, validação cruzada, experimentos com usuários reais, benchmarking

Trabalhos futuros: Sugestões de melhorias

- Integração de outras fontes de dados.
- Dados contextuais: inclusão de informações temporais, geográficas e demográficas na geração de recomendações.
- Feedback implícito x Feedback explícito
- Utilização de datasets de literatura nacional.
- Entre outros.

Obrigado!

- Contato:

10440929@mackenzista.com.br

10441014@mackenzista.com.br

10433408@mackenzista.com.br

- Link do GitHub:
  https://github.com/Wesrsant/projetoaplicadomackenzie_III